The Data Wrangle and Analysis Project

By **Youssef Senousy**

The main goal of this project is to make a good wrangle and analysis of the Twitter account @WeRateDogs from different data resources and obtain at least 8 quality issues and at least 2 tidiness issues and fix them to produce effective analysis.
We have three types of datasets:

1. Enhanced Twitter Archive Dataset
2. Predicted Images Dataset
3. Retrieved Data from Twitter API.

Project Steps:

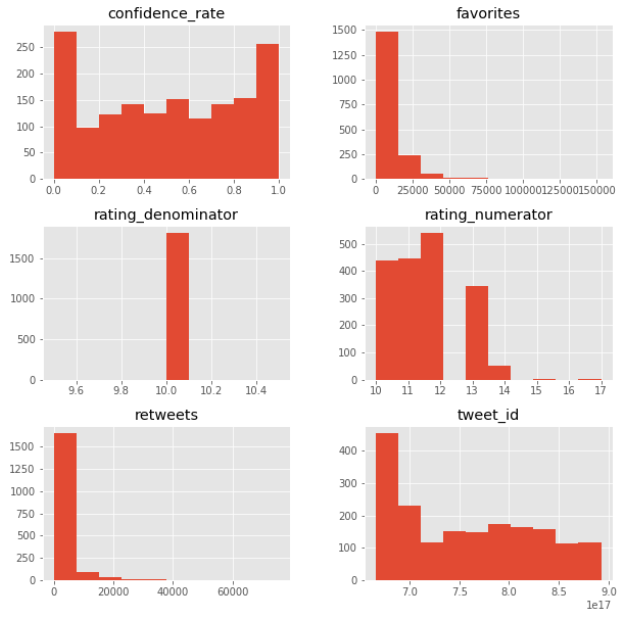
a) Data wrangling which consists of three main steps:

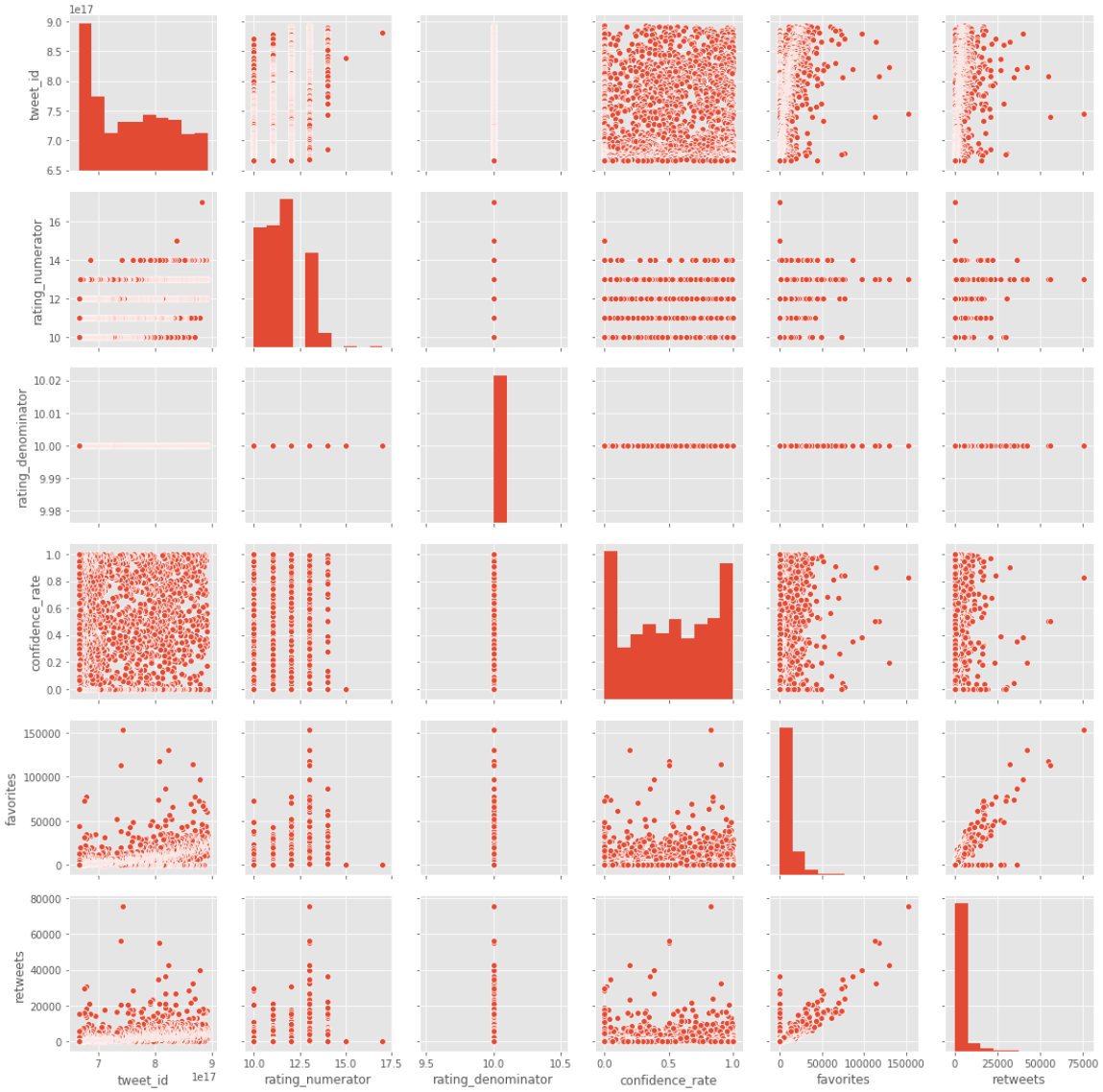
1. Gathering Datasets
2. Assessing Datasets
3. Cleaning Datasets

b) Storing the datasets after the steps mentioned above to obtain report and visualization of data.

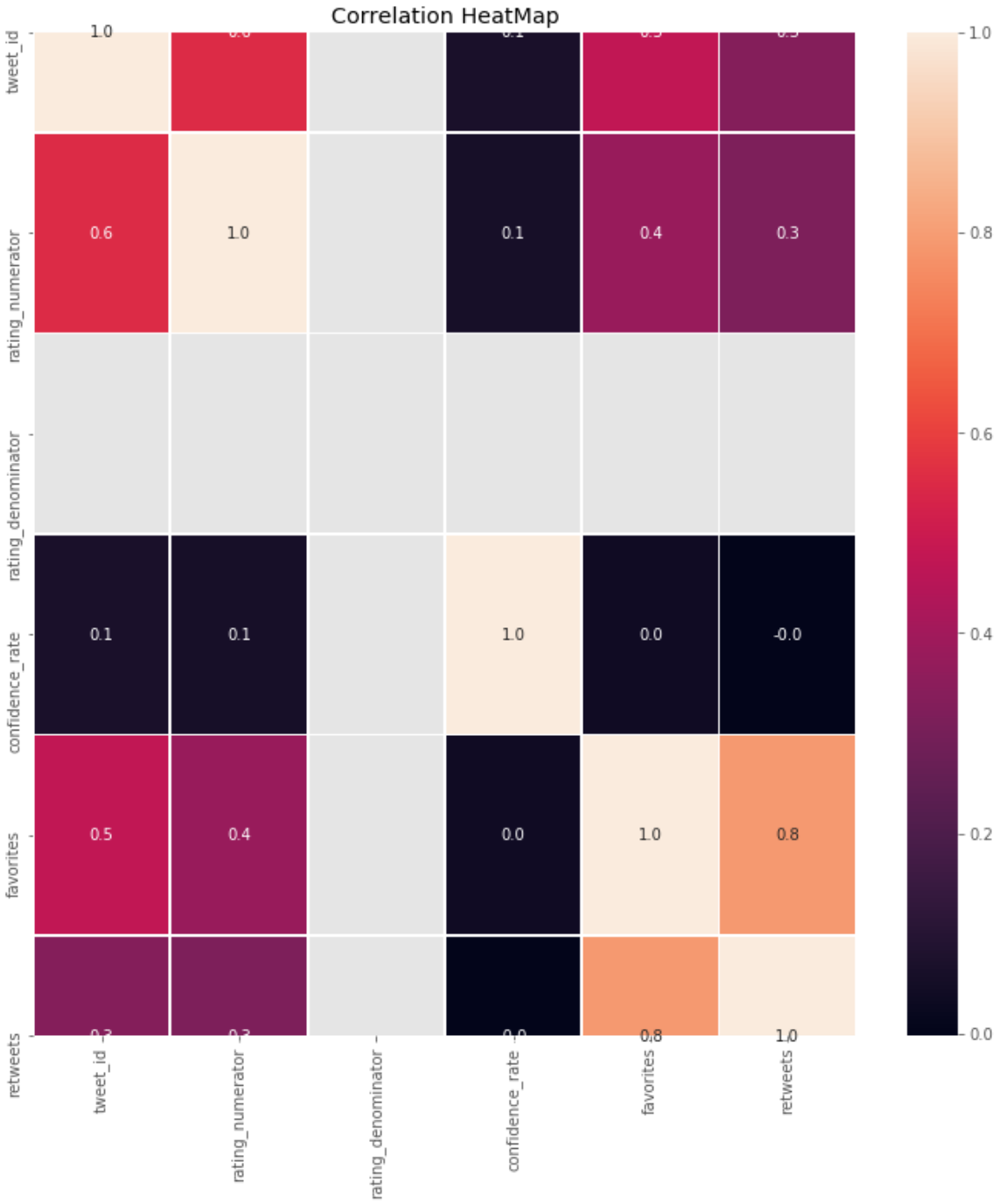
After finishing the steps mentioned above, we found some insights from the final dataset by using some visualizations.

A) Histograms:



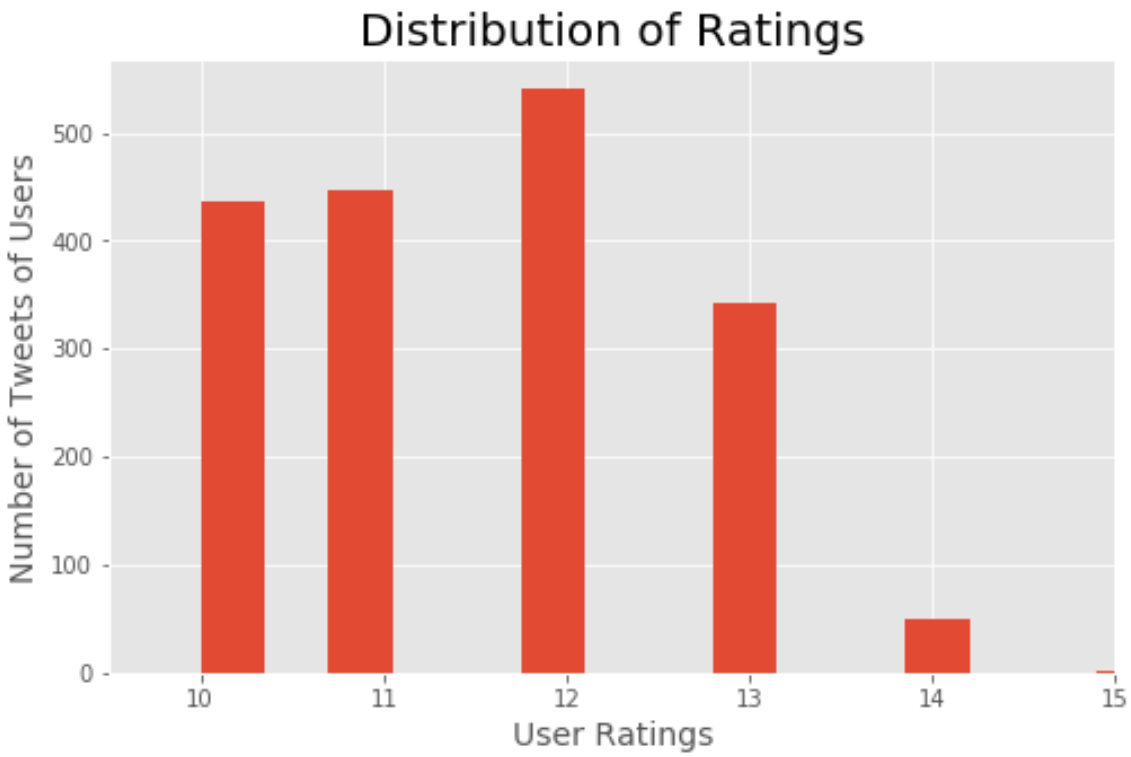

tweet_id
rating_numerator
rating_denominator
confidence_rate
favorites
retweets
1e17

B) Correlation Heat map

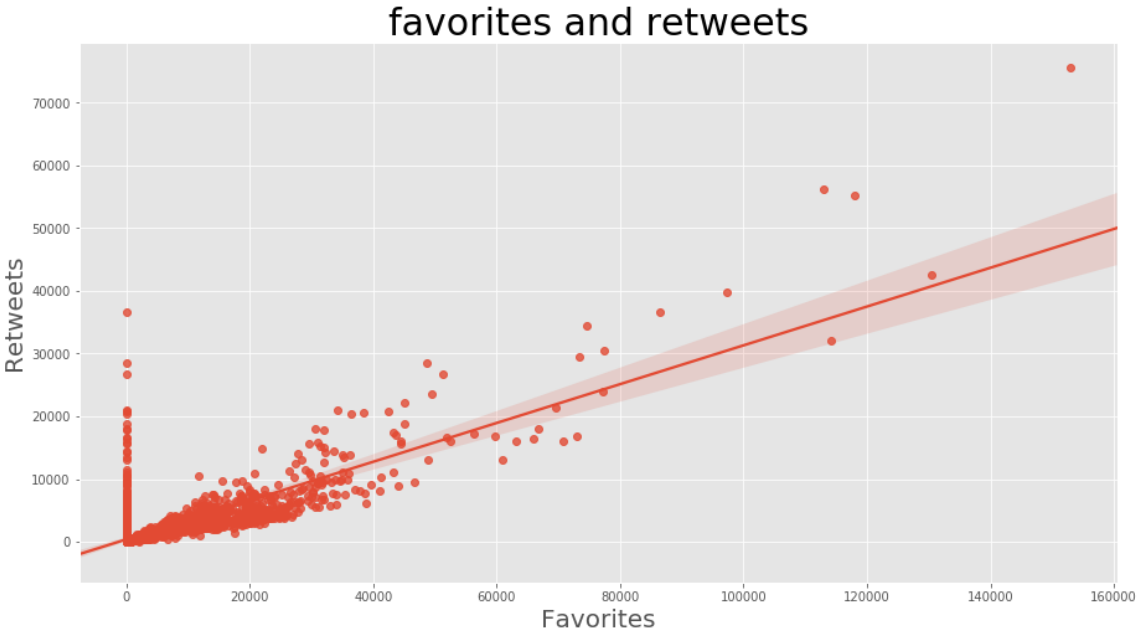


Features Analysis

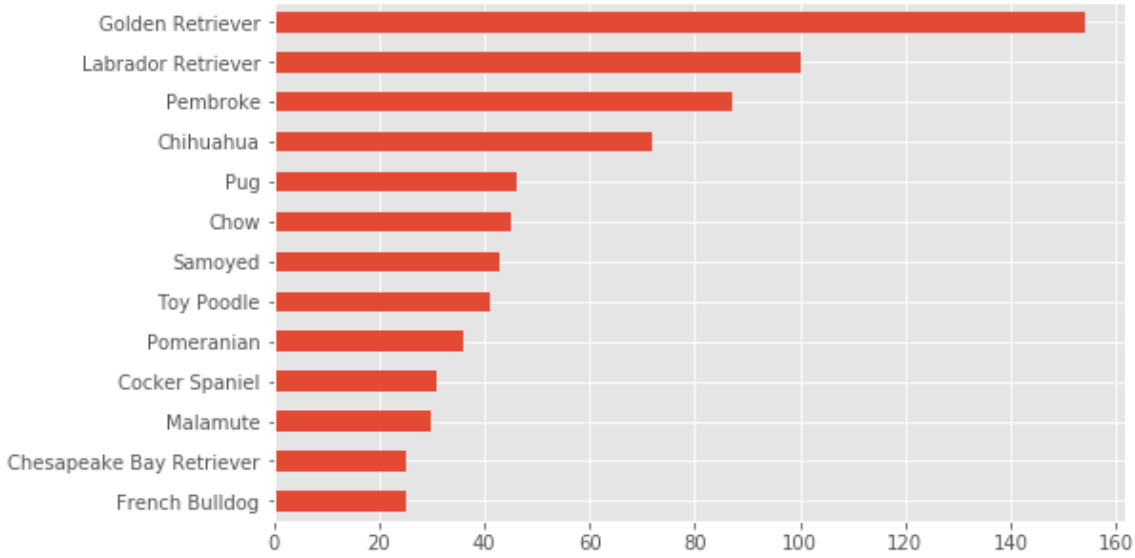
Distribution of Ratings with Number of Tweets



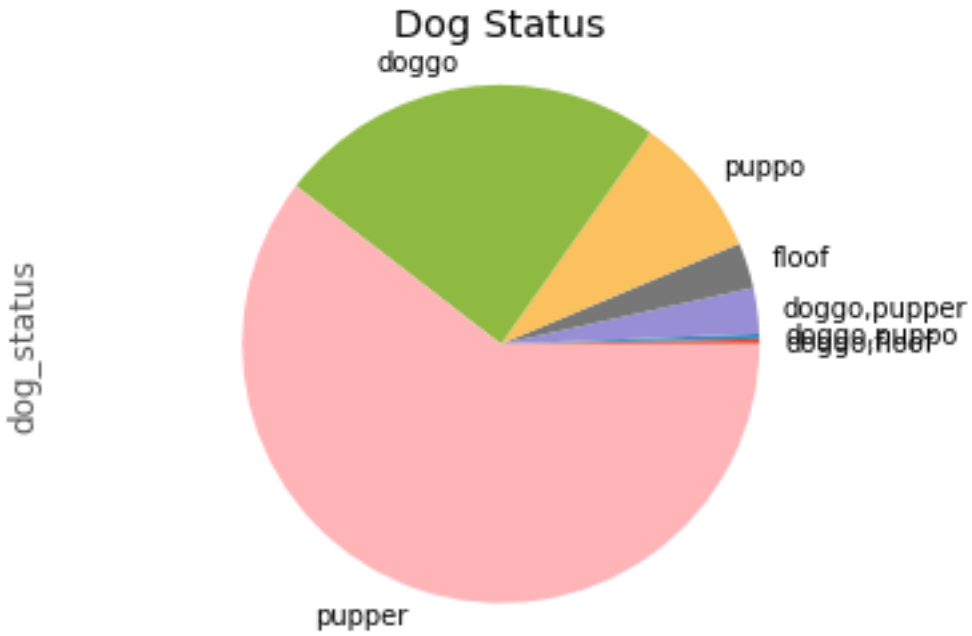
From correlation heat map we can see the strong correlation between favorite_count and retweet_count



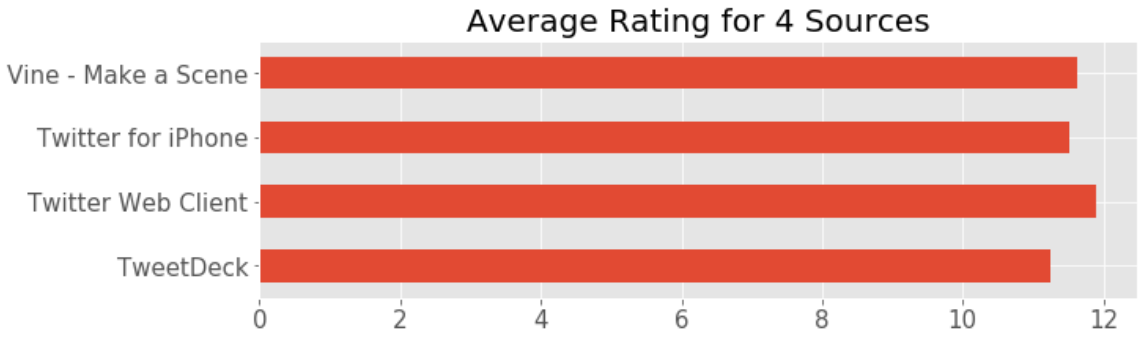
The most common breed types from the dataset



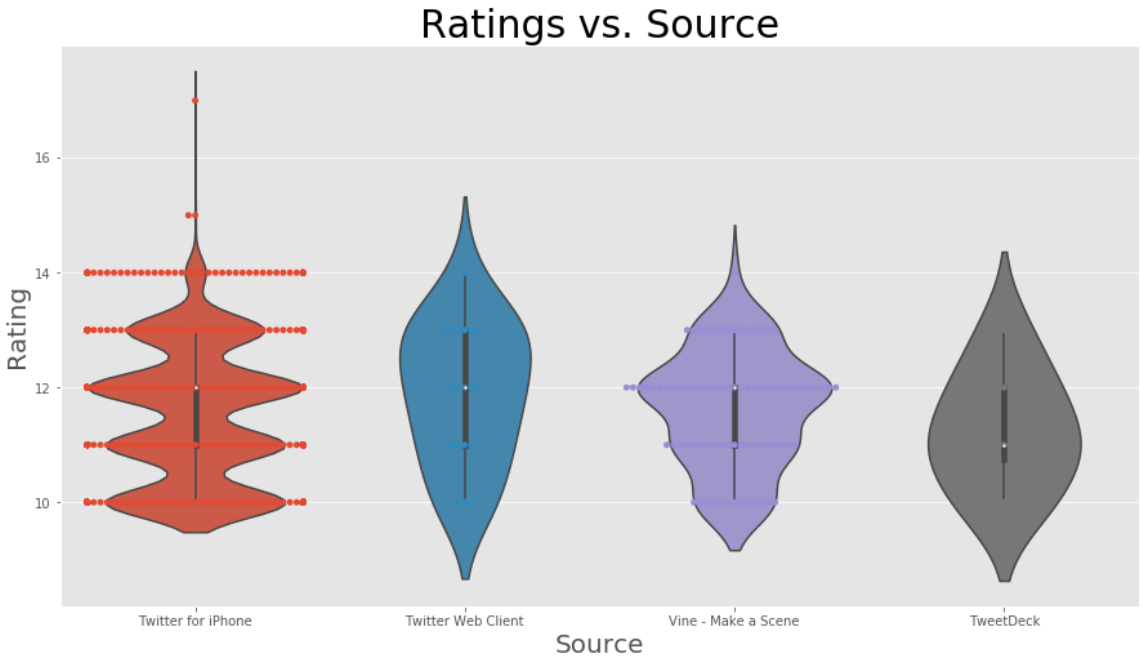
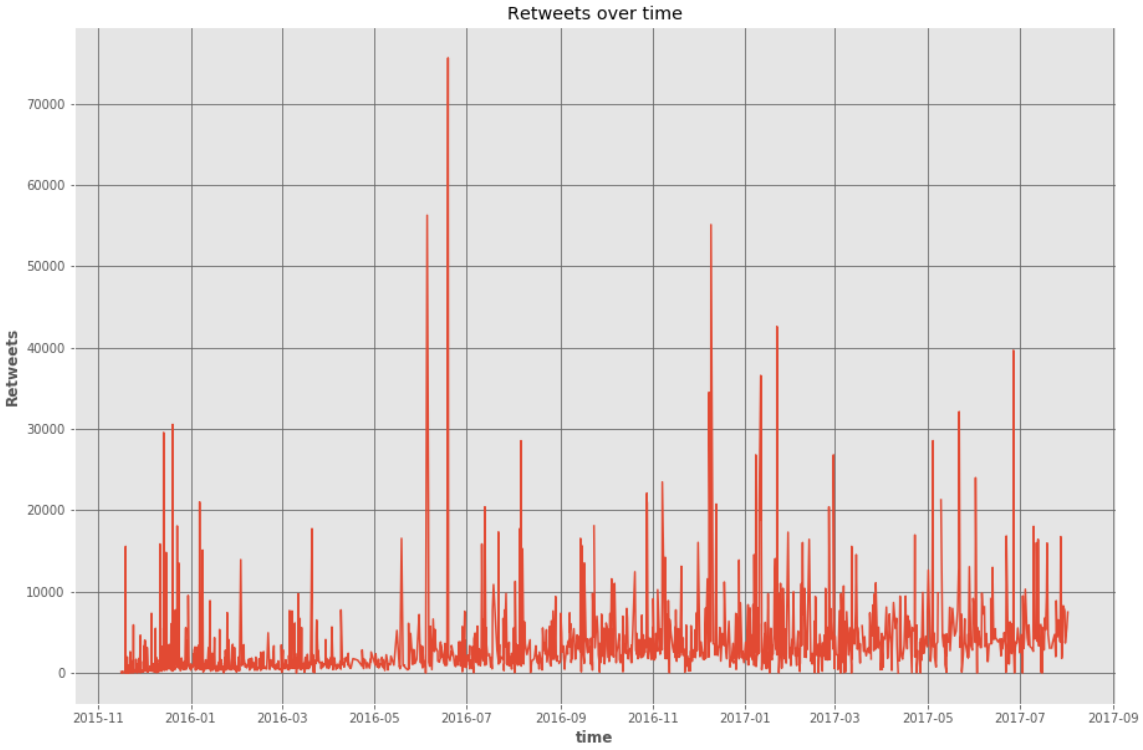
The Pie Chart of the dog status



The average source of the tweets distributions



The relation between ratings and its sources.



The tweets of users distributed by time